J. Mod. Transport. (2016) 24(3):224–235
DOI 10.1007/s40534-016-0110-x

www.springer.com/40534

Key factors contributing to crash severity at highway-rail grade crossings

Wei Fan[1] · Linfeng Gong[1] · Edward Matt Washing[1] · Miao Yu[1] · Elias Haile[2]

Received: 17 February 2016 / Revised: 18 May 2016 / Accepted: 25 May 2016 / Published online: 1 July 2016

Abstract The purpose of this paper is to develop and compare the preferred multinomial logit (MNL) and ordered logit (ORL) model in identifying factors that are important in making an injury severity difference and exploring the impact of such explanatory variables on three different severity levels of vehicle-related crashes at highway-rail grade crossings (HRGCs) in the United States. Vehicle-rail crash data on USDOT highway-rail crossing inventory and public crossing sites from 2005 to 2012 are used in this study. Preferred MNL and ORL models are developed and marginal effects are also calculated and compared. A majority of the variables have shown similar effects on the probability of the three different severity levels in both models. In addition, based on the Akaike information criterion, it is found that the MNL model is better than the ORL model in predicting the vehicle crash severity levels on HRGCs in this study. Therefore, the researchers recommend the use of MNL model in predicting severity levels of vehicle-rail crashes on HRGCs.

✉ Wei Fan
wfan7@uncc.edu

Linfeng Gong
lgong@uncc.edu

Edward Matt Washing
ewashin5@uncc.edu

Miao Yu
myu6@uncc.edu

Elias Haile
gurjimeron@gmail.com

[1] Department of Civil and Environmental Engineering, The University of North Carolina at Charlotte, EPIC Building, Room 3261, 9201 University City Boulevard, Charlotte, NC 28223, USA

[2] 1721 Market St. # 102., Oakland, CA 94607, USA

Keywords Vehicle crashes · Severity · Highway-rail grade crossings · Multinomial logit model · Ordered logit model

1 Introduction

Fatality resulting from motor vehicle crashes is the fifth leading cause of death in the United States. Data from the National Highway Traffic Safety Administration indicates that since 1949, more than 30,000 (40,000 average) fatalities result from motor vehicle crashes every year. However, the current trend shows this number is declining. For example, a 1.9 % decrease in crash-related fatalities was observed in 2011 as compared to 2010. Nonetheless, crash-related injuries are still large in number. In 2011, estimated 2.22 million people were injured in motor vehicle traffic crashes and 2.24 million in 2010 [1]. Fatal crashes on highway-rail grade crossings (HRGCs) contributed to 261 deaths in 2010 and 251 in 2011 [2].

HRGCs are conflict points between highway users and rail equipment (e.g., locomotive, freight car, caboose, or service equipment car operated by a railway company), which has contributed to a considerable amount of crashes in the U.S. history. Although the trend of highway user crashes with rail equipment is showing a decrease in numbers, much has yet to be done to improve the safety of HRGCs. Unlike highway traffic crashes, a significantly high percentage of vehicle-rail crashes lead to fatality and injury to vehicle users. For example, data in the past 8 years (2005–2012) indicate that 8.55 % of vehicle-rail crashes were fatal and 26.68 % resulted in injury [2]. However, in the case of highway traffic crashes, the percentage of fatal crashes is not more than 2 % [1].

Despite the fact that highway user-rail crashes have a significant effect on highway user safety, the subject (of examining the injury severity levels in such crashes) still receives little attention. An understanding of the factors contributing to the levels of injury severity is an important step toward making the transportation system safer and more attractive. Responsible jurisdictions may use the results of this research to derive road user safety measures and policies.

One of the most important tasks in improving road safety is to uncover influential factors and then to develop countermeasures. The relationship between the injury severity of traffic crashes and factors such as driver and passenger characteristics, pedestrian age and gender, vehicle type, environmental conditions, and traffic and geometric conditions has attracted much attention. A better understanding of this relationship is necessary and very important for improving facility design so that accidents can be reduced. It is important to note that reducing crash frequency and/or reducing crash–injury severity may necessitate different strategic approaches. The development of effective countermeasures requires a thorough understanding of the factors that affect the likelihood of a crash occurring or, given that a crash has occurred, the characteristics that may mitigate or exacerbate the degree of injury sustained by crash-involved road users. To gain such an understanding, safety researchers have applied a wide variety of methodological techniques over the years.

Numerous studies have applied statistical models for crash–injury severity. Among them, the unordered logit, ordered logit models, and their variations are the most often used models. Savolainen et al. [3] briefly discussed and summarized a wide range of methodological tools applied to study the impact of various factors on motor vehicle crash–injury severity. As presented in the paper, ordered logit and probit, multinomial logit, binary logit and binary probit, and nested logit are some of the frequently used statistical methodologies. In particular, logistic regression has been widely applied to model crash severity levels. Variables such as elements of geometric design, traffic operational measures, and environmental conditions are considered as independent variables to estimate the severity. In particular, it is important to note that modeling ordinal outcome-dependent variable using nominal variable will lead to loss of efficiency as a result of ignoring information. In the reverse, modeling nominal variable using ordinal variable will give biased or sometimes irrational estimates [4].

As discussed, crashes occurring at HRGCs have a significant effect on highway user safety and the importance of conducting research in such areas is evident. However, this subject receives less attention and little research efforts have been made in this particular area (except [5]) in which a multinomial logit model was developed to analyze the severity of vehicle crashes at HRGCs). As such, the objective of this research is to apply and compare the multinomial logit and ordered logit models to explore the impacts of various factors contributing to different levels of crash severity to vehicle users as a result of vehicle-rail crashes on HRGCs.

The remainder of this paper is organized as follows. First, Sect. 2 presents a literature review of existing studies regarding vehicle crash severity modeling. Then, Sect. 3 describes the multinomial logit (MNL) and ordered logit (ORL) modeling methodology. Section 4 discusses the data assembly and analysis of the research. Section 5 presents numerical results and discussion. Finally, conclusions and recommendations are made in Sect. 6.

2 Literature review

Several studies have been conducted to model crash severity and investigate the impacts of various factors involved in the crashes. Mercier et al. [6] conducted a study and tested the hypothesis that older drivers and passengers would suffer more severe injuries than younger adults in the presence of broadside and angle collisions of automobiles on rural highways. Logistic modeling, Hierarchical Regression Analysis, and Principal Components Regression were analysis tools applied. Injury severity levels, fatal, major, and minor were considered as dependent categorical variables. Some of the independent variables which were considered included occupant age, occupant position relative to point of impact and protection. According to the study, age is reported as a significant predictor of injury severity and is slightly higher for females than males. It was also identified that the use of lap and shoulder restrains, reduces injury severity, and is less certain for females. For females only, air bags deployed were reported as significant injury severity predictors.

By using sequential binary logistic regression, Dissanayake and Lu [7] modeled crash severity for single-vehicle fixed object crashes involving young drivers. The five crash severity categories which were considered including no injury, possible injury, noncapacitating injury, incapacitating injury, and fatal. As reported in the study, factors such as alcohol or drug influence, ejection in the crash, point of impact, rural crash locations, existence of curve or grade at the crash location, and speed of vehicle significantly increased the probability of more severe crashes. On the other hand, restraint device usage and drivers being of male gender were reported to reduce the chance of high severity crashes. It was also indicated that factors such as weather condition, residence location, and physical condition had no significant relation in the model.

Duncan et al. [8] conducted a study to investigate car occupant injury severity in two-vehicle passenger car-truck rear-end collisions by an ordered probit model. As reported in the study, factors such as darkness, large speed differentials, high speed limits, grades, being in a car struck to the rear, driving while drunk, and being female increased the passenger vehicle occupant injury severity. On the other hand, factors such as snowy or icy roads, being in a child restraint, and congested roads decreased the severity level. It was also indicated that interaction effects of cars being struck to the rear with large speed differentials and car rollovers were significant.

Donnell and Mason [9] conducted a study and developed median-related crash severity models. Three crash severity classes, fatal, injury, and property damage only (PDO) were considered as independent variable outcome. Both ordinal and nominal response logistic regression models were developed in the study. As indicated in the report, the ordinal response model gave more attractive results for cross-median crashes. On the other hand, the nominal response model gave better result for median-barrier crashes. Furthermore, variables such as highway surface conditions, use of drugs or alcohol, presence of an interchange entrance ramp, horizontal alignment, crash type, and average daily traffic volume were reported to have effect on crash severity.

By using paired comparison analysis and ordered probit model, Renski et al. [10] conducted a study to test the hypothesis that a speed limit increase will result in an increase in driving speed and produce higher crash severity. The study was focused on single-vehicle crashes on interstate roadways in North Carolina. As reported in the study, increasing speed limits from 89 to 97 km/h and from 89 to 105 km/h increased the probability of sustaining minor and noncapacitating injuries. However, the study indicated that increasing speed limits from 105 to 113 km/h did not show significant effect on crash severity [10].

Huang et al. [11] investigated effects of road diets in which four-lane undivided roads were converted into three lanes. Twelve road diets and 25 comparison sites in California and Washington cities were considered in the study. A "yoked comparison" study was applied to a "before" and "after" analysis and it was reported that road-diet crashes were observed to be lower by 6 percent than that of the comparison sites before road-diet countermeasures were made. Khattak [12] conducted a study that investigated the effect of vehicle technologies on crash–injury severity. The North Carolina 1994–1995 HSIS crash data were used for the analysis. Three separate ordered probit models were developed for the three drivers, Driver 1 (leading), Driver 2 (striking), and Driver 3 (striking in a three-vehicle crash). As indicated in the study, in a two-vehicle rear-end collision the leading driver is more likely to be injured, whereas in a three-vehicle collision, the driver in the middle is more likely to be injured. It was also stated that being in a newer vehicle protects the driver in rear-end collisions. Moreover, the study showed the benefit of technological improvements on driver safety.

Mercier et al. [13] performed a study and tested the hypothesis that older drivers and passengers would suffer more severe injuries than younger adults in the presence of head-on collisions of automobiles on rural highways. Logistic modeling, Hierarchical Regression Analysis, and Principal Components Regression were applied. Injury severity levels fatal, major, and minor were considered as dependent categorical variable. The independent variables considered included, among others, occupant age, occupant position relative to point of impact, and level of protection. As stated in the study, age was an important factor in predicting injury severity for both men and women. The study concluded that older drivers and passengers experienced more severe injuries than any of other age groups. The use of lap and shoulder devices was reported to be more important for men than women while the reverse is true for deployed air bags.

Chira-Chavala et al. [14] investigated the characteristics and probable causes of light rail transit system accidents and developed a crash severity model for the Santa Clara County Transit Agency. A binary logit model was applied to predict the probability of injury accident as a function of explanatory variables such as speeds before collision of light rail vehicles and motor vehicles, movement of the motor vehicle before collision, etc. As reported in the study, left-turn vehicle movements, higher speeds of the motor vehicle, or the LRV and accident occurring during peak hours increased the probability of injury accidents.

Chen and Jovanis [15] developed and tested the variable-selection procedure that avoids problems occurring due to the presence of the large number of potential factors, the complex nature of crash causes and outcomes, and the large number of categories in crash severity modeling. Bus-involved crash data for Freeway 1 in Taiwan from 1985 through 1993 were used. The procedure consisted of the χ^2 automated interaction detection (CHAID) method to collapse categories. Pearson χ^2 test was used to assess the relationship between dependent and independent variables, and log-linear modeling techniques. As indicated in the study, the log-linear model showed that late-night or early-morning driving increased the risk of severe injury crashes for bus drivers. It was also stated that bus crashes involving a large truck or tractor-trailers increased the risk of severe injury crashes.

By using an ordered probit model, Khattak et al. [16], explored factors contributing to more severe older driver (age of 65 and above) crash–injury severity by analyzing 1990–1999 crash data from Iowa. According to the study,

older male drivers are more prone to injury as compared to older female drivers. It was stated that older drivers under the influence of alcohol experienced more severe injuries. It was also indicated that older driver injuries involving farm vehicles are more severe as compared to other vehicle types. Xie et al. [17] conducted a study that demonstrated application of a Bayesian ordered probit model in drivers' injury severity analysis. In the Bayesian probit model, prior distributions such as means and variances were included (reflecting the analysts' prior knowledge about the data). Comparisons were made between Bayesian ordered probit and conventional ordered probit models. As reported in the study, for large data size, model fitting results obtained from the Bayesian and the conventional probit model have no significant differences. It was also reported that for small sample size, a Bayesian probit model produced parameter estimates with better prediction performance than the conventional ordered probit model. Most recently, Zhao and Khattak [18] modeled motor vehicle drivers' injuries in train-motor vehicle.

The purpose of this study is to analyze severity of vehicle crashes on HRGCs and make comparison between ordered and unordered logistic regression models in predicting vehicle crash severities at HRGCs. MNL and ORL models are developed to model the impact of various factors which include vehicle driver characteristics, environmental factors, railroad crossing characteristics, highway characteristics, land use type, and more, using the same dataset. Since the coefficients cannot be compared directly, marginal effects/values are computed for both models. The three levels of responses considered are fatality, injury, and no injury. The SAS PROC LOGISTIC procedure is used to develop the models.

3 Modeling methodology

3.1 MNL model

The MNL model formulation is well discussed by Long [4]. If y is the response variable with J nominal (i.e., categorical) outcomes (which takes on one of a limited number of possible values), then the assumption of the multinomial logit model is that category 1 through J are not ordered (i.e., not arranged in an increasing or decreasing order). Also, let $\Pr(y = m|x)$ be the probability of observing outcome m given the independent variable x. The model for y is constructed as follows:

- Assume that $\Pr(y = m|x)$ is a linear combination $\boldsymbol{x\beta_m}$. The vector $\boldsymbol{\beta_m} = (\beta_{0m}\ldots\beta_{km}\ldots\beta_{km})$ contains the intercept β_{0m} and coefficients of β_{km} for the effects of x_k on outcome m.
- To ensure nonnegativity for the probabilities, the exponential of $\boldsymbol{x\beta_m}$ is used.
- For the probabilities to sum to 1, divide $\exp(\boldsymbol{x\beta_m})$ by $\sum_{j=1}^{J}\exp(x_i\beta_j)$.

$$\Pr(y_i = m|x_i) = \frac{\exp(x_i\boldsymbol{\beta_m})}{\sum_{j=1}^{J}\exp(x_i\beta_j)}. \quad (1)$$

Although the probability sum is 1, the set of parameters that generate the probabilities is not identified since more than one set of parameters can generate the same probabilities. In order to identify the set of parameters that generate the probabilities, a constant must be imposed. By imposing one of the parameter estimates to be equal to zero (assume $\beta_1 = 0$), the model can be written as follows:

$$\Pr(y_i = 1|x_i) = \frac{1}{1 + \sum_{j=2}^{J}\exp(x_i\beta_j)}. \quad (2)$$

$$\Pr(y_i = m|x_i) = \frac{\exp(x_i\boldsymbol{\beta_m})}{1 + \sum_{j=2}^{J}\exp(x_i\beta_j)} \text{ for } m > 1. \quad (3)$$

The parameter estimates are determined using maximum likelihood estimation. If the observations are independent, the likelihood Eq. (4) is given by

$$L(\beta_2, \ldots.\beta_J|y, x) = \prod_{i=1}^{N} P_i, \quad (4)$$

where P_i is the probability of observing whether values of y was actually observed for the ith observation. Combining the Eq. (1) with this Eq. (4) in place of P_i the likelihood Eq. (5) can be written as

$$L(\beta_2, \ldots\beta_J|y, x) = \prod_{m=1}^{J}\prod_{y_{i=m}}\frac{\exp(x_i\boldsymbol{\beta_m})}{\sum_{j=1}^{J}\exp(x_i\beta_j)}, \quad (5)$$

where $\prod_{y_{i=m}}$ is the product over all cases for which y_i is equal to m. Taking logs, we may obtain the log-likelihood function which can be maximized with numerical methods to estimate the β's.

The overall model fitness can be compared using the model's log-likelihood at convergence with the log-likelihood of a naive model (model with all coefficients set to zero which is equivalent to assigning equal probability for all outcomes). It is also possible to compare a model with only alternative constants (assigning probability to outcomes equal to the observed share of the outcomes in the dataset).

$$\rho^2 = 1 - \frac{\mathrm{LL}(\beta)}{\mathrm{LL}(0)}, \tag{6}$$

where LL(β) represents the log-likelihood at model convergence, LL(0) represents the log-likelihood of a naïve model (without information). The ρ^2 goes from 0 (for no improvement in the log-likelihood) to 1 for a perfect fit. A value for ρ^2 larger than 0.1 indicates meaningful improvement [4].

The marginal effect or partial change can be determined by taking derivative of Eq. (1) with respect to x_k as described in the following Eq. (7).

$$\frac{\partial \Pr(y = m|x)}{\partial x_k} = \Pr(y = m|x)\left[\beta_{km} - \sum_{j=1}^{J} \beta_{kj} \Pr(y = j|x)\right]. \tag{7}$$

Marginal effect is the slope of the curve relating xk to $\Pr(y = m|x)$, holding other variables constant. Variables are held at their means, possibly with dummy variables at 0 or 1. Although the computation of the change in the probability is important to interpret the effects of the MNL model, there is limitation in that it measures the discrete change which does not indicate the changes among the dependent outcome due to infinitely small changes in independent variables [4].

Odds ratio can also be used in the interpretation of the developed model. The odds ratio is defined as the ratio of the odds of those with the risk factor to the odds for those without the risk factor. Generally, the odds ratio associated with a one-unit increase in the risk factor can be computed by the exponential function of the regression coefficient of that risk factor [19].

3.2 Ordered logit (ORL) model

When the absolute distance between categories of a variable is unknown, yet there is a clear ordering of the categories, the variable is considered ordinal. The ordered response logistic regression formulation is presented as discussed by Long [4]. An ordinal logistic regression model is derived from a measurement model in which a latent variable y^* is mapped to an observed variable y. These variables are related according to the following equation:

$$y_i = m \text{ if } \tau_{m-1} \le y_i^* < \tau_m \text{ for } m = 1 \textit{ to } J. \tag{8}$$

The τ's are cutpoints on the measurement scale that are used to distinguish the ordinal categories. In the case of crash severity models, the ordinal response categories are fatality, injury, and no injury crashes. Category 1 (fatal) is defined by the open-ended interval on the lower end of the measurement scale; Category 3 (no injury), is defined as the portion of the scale above cut point τ_2 and Category 2 (injury) is the portion between the two cutpoints. Note that the crash severities are divided into 3 categories instead of 5 levels due to the data availability issues in this paper. In other words, if 5 levels are used, all the 3 injury categories (except fatal and no injury categories) will each involve only a very limited amount of data for modeling, which may cause some data underrepresentation issues. As such, to avoid such issues, the crash severities are classified into only 3 categories, which are used for modeling in this paper.

The observed y is related to y^* according to the measurement model:

$$y_i = \begin{cases} 1 = \text{Fatal} & \text{if } \tau_o = -\infty \le y_i^* < \tau_1 \\ 2 = \text{Injury} & \text{if } \tau_1 \le y_i^* < \tau_2 \\ 3 = \text{No injury} & \text{if } \tau_2 \le y_i^* < \tau_3 = \infty \end{cases}. \tag{9}$$

The regression equation used for an ordinal response is $y_i^* = \boldsymbol{x_i\beta} + \varepsilon_i$, where $\boldsymbol{x_i}$ is a row vector (with 1 in the first column for the intercept), $\boldsymbol{\beta}$ is a column vector of structural coefficients (with the first element being the intercept βo), and ε_i is an error term.

Maximum likelihood (ML) estimation can be used to estimate the regression of y^* on x. In order to use ML, assumption of a specific type of error (ε) distribution is required. For the ordered logit model, the error term has a logistic distribution with mean zero and a variance of $\pi^2/3$. The probability density function (p.d.f) of the logistic distribution is given as shown in Eq. (10).

$$\lambda(\varepsilon) = \frac{\exp(\varepsilon)}{[1 + \exp(\varepsilon)]^2}. \tag{10}$$

The cumulative distribution function (c.d.f) of the logistic distribution is given as shown in Eq. (11):

$$\Lambda(\varepsilon) = \frac{\exp(\varepsilon)}{1 + \exp(\varepsilon)}. \tag{11}$$

After specification of the error term, the probabilities of observing values of y given x can be computed. The probability of any observed outcome y = m given x is the difference between the c.d.f evaluated at these values:

$$\Pr(y_i = m|X_i) = F(\tau_m - \boldsymbol{x_i\beta}) - F(\tau_{m-1} - \boldsymbol{x_i\beta}). \tag{12}$$

To estimate the model, let $\boldsymbol{\beta}$ be the vector with parameters from the structural model, with the intercept βo in the first row and let τ be the vector containing the threshold parameters. Either βo or τ_1 is constrained to 0 to identify the model. Program such as SAS's LOGISTIC procedure assumes βo and estimates τ_1. From Eq. (12), the following can be obtained:

$$\Pr(y_i = m|X_i, \boldsymbol{\beta}, \tau) = F(\tau_m - \boldsymbol{x_i\beta}) - F(\tau_{m-1} - \boldsymbol{x_i\beta}). \quad (13)$$

The probability of observing whatever value of y was actually observed for the ith observation is

$$P_i = \begin{cases} \Pr(y_i = 1|x_i, \boldsymbol{\beta}, \tau) \text{ if } y = 1 \\ \quad . \\ \quad . \\ \quad . \\ \Pr(y_i = m|x_i, \boldsymbol{\beta}, \tau) \text{ if } y = m\,. \\ \quad . \\ \quad . \\ \quad . \\ \Pr(y_i = J|x_i, \boldsymbol{\beta}, \tau) \text{ if } y = J \end{cases} \quad (14)$$

If the observations are independent, the likelihood equation over the population of N observations is

$$L(\boldsymbol{\beta}, \tau|y, X) = \prod_{i=1}^{N} P_i. \quad (15)$$

Combining Eqs. (13), (14), and (15),

$$\begin{aligned} L(\beta, \tau|y, X) &= \prod_{j=1}^{J} \prod_{y_i=j} \Pr(y_i = j|x_i, \boldsymbol{\beta}, \tau) \\ &= \prod_{j=1}^{J} \prod_{y_i=j} [F(\tau_j - \boldsymbol{x_i\beta}) - F(\tau_{j-1} - \boldsymbol{x_i\beta})]. \end{aligned} \quad (16)$$

Here, $\prod_{y_i=j}$ indicates multiplying over all cases, where y is observed to equal j. Taking logs, the log-likelihood can be written as follows:

$$\ln L(\beta, \tau|y, X) = \sum_{j=1}^{J} \sum_{y_i=j} \ln[F(\tau_j - \boldsymbol{x_i\beta}) - F(\tau_{j-1} - \boldsymbol{x_i\beta})]. \quad (17)$$

Model estimation involves maximizing Eq. (17) using numerical methods to estimate the τ's and the β's.

A measure of the model goodness of fit (ρ^2) can be calculated as

$$\rho^2 = 1 - \left[\frac{\ln L_b}{\ln L_o}\right], \quad (18)$$

where $\ln L_b$ is the log-likelihood at convergence and $\ln L_o$ is the restricted log-likelihood. The ρ^2 measure is bound by zero and one. Values of ρ^2 closer to one indicate better fit of the model.

Interpretation of ordinal response variables can be performed according to odds ratios. In this paper, the proportional odds model is used to interpret odds ratios for cumulative probabilities.

The cumulative probability that the outcome is less than or equal to m is

$$\Pr(y \le m|x) = \sum_{j=1}^{m} \Pr(y = j|x) \text{for } m = 1, \ldots, J-1. \quad (19)$$

The odds that an outcome is m or less versus greater than m given a set of explanatory variables x are

$$\Omega_m(x) = \frac{\Pr(y \le m|x)}{1 - Pr(y \le m|x)} = \frac{\Pr(y \le m|x)}{\Pr(y > m|x)} = \exp(\tau_m - \boldsymbol{x\beta}). \quad (20)$$

Taking the log results in the logit equation,

$$\ln\Omega_m(x) = \tau_m - x\beta. \quad (21)$$

The marginal effects of variables x on the underlying crash severity propensity can be evaluated by taking the partial derivative of Eq. (11) with respect to x_k, resulting in

$$\frac{\partial \Pr(y = m|x)}{\partial x_k} = \frac{\partial F(\tau_m - \boldsymbol{x\beta})}{\partial x_k} - \frac{\partial F(\tau_{m-1} - \boldsymbol{x\beta})}{\partial x_k}, \quad (22)$$

or

$$\frac{\partial \Pr(y = m|x)}{\partial x_k} = \beta_k[f(\tau_{m-1} - \boldsymbol{x\beta}) - f(\tau_m - \boldsymbol{x\beta})]. \quad (23)$$

The marginal effect is the slope of the curve relating x_k to $\Pr(y = m|\boldsymbol{x})$, holding all other variables constant and is usually computed at the mean values of all variables. For a dummy independent variable, the derivative while treating it as a continuous variable provides an approximation.

4 Data assembly and model description

Vehicle-rail crash data on the USDOT public crossing sites from 2005 to 2012 are used in this study. In order to acquire more explanatory variables, the USDOT highway-rail crossing inventory is also included. The crash data and the crossing inventory data are merged based on the USDOT identification number. The SAS PROC SQL is used to merge and clean the data. After the data merging and cleaning process, a total of 7,414 records are obtained and used in the modeling stage from the Federal Railroad Administration (FRA) database. The data used to create the dataset are obtained from the FRA [2].

Table 1 presents the descriptive statistics of some of the variables from such HRGC crash and inventory data. As shown, the distribution of vehicle-rail crash severity is 6.80 %, 26.63 %, and 66.58 % for fatal, injury, and no injury, respectively. This distribution of crash severity indicates around 33.43 % of vehicle crashes at HRGC sites lead to fatality or injury, in which the figures are much higher as compared to those of multi-vehicle crashes in highway traffic. The majority (78.64 %) of vehicle-rail crashes at HRGC sites occurred when the rail equipment

Table 1 Descriptive statistics of variables from HRGC crash and inventory data

Variable	Category	Frequency	Percent
Crash characteristics			
INJURY (crash severity level)	3 = fatal crashes	504	6.80
	2 = injury crashes	1974	26.63
	1 = no Injury crashes	4936	66.58
TYPACC (type of accident)	1 = train struck vehicle	5830	78.64
	2 = vehicle struck train	1584	21.36
Vehicle characteristics			
TYPVEH (type of vehicle)	1 = auto	3,936	53.09
	2 = truck	542	7.31
	3 = truck-trailer	1,298	17.51
	4 = pick-up truck	1,317	17.76
	5 = van	306	4.13
	6 = bus	10	0.13
	7 = school bus	5	0.07
VEHSPD (vehicle speed)	1 = <40 km/h	6,312	85.14
	2 = 40–72 km/h	830	11.20
	3 =>72 km/h	272	3.67
AADT (average annual daily traffic)	1 = <10,000	6,525	88.01
	2 = 10,000–20,000	602	8.12
	3 = 20,000–30,000	177	2.39
	4 =>30,000	110	1.48
Train characteristics			
TRNSPD (train speed)	1 = <40 km/h	2,999	40.45
	2 = 40-72 km/h	2,549	34.38
	3 =>72 km/h	1,866	25.17
Vehicle driver characteristics			
DRVAGE (driver age)	1 = <25 years	1,186	16.00
	2 = 25–60 years	3,978	53.66
	3 =>60 years	1,029	13.88
	Missing	1,221	16.47
DRIVGEN (vehicle driver gender)	1 = male	5,645	76.14
	2 = female	1,769	23.86
Highway characteristics			
HWYPVED (highway surface type)	1 = paved	6,042	81.49
	2 = unpaved	1,372	18.51
HWYSGNL (highway signal)	1 = not present	7,215	97.32
	2 = present	199	2.68
TRAFICLN (no. of traffic lane)	1 = 1 lane	644	8.69
	2 = 2 lanes	5,560	74.99
	3 = 3 lanes	87	1.17
	4 = 4 lanes	872	11.76
	5 = ≥5 lanes	251	3.39
Environmental characteristics			
DEVELTYP (development area type)	1 = open space	2,400	32.37
	2 = residential	1,595	21.51
	3 = commercial	2,083	28.1
	4 = industrial	1,226	16.54
	5 = institutional	110	1.48
WEATHER (weather condition)	1 = clear	5,265	71.01
	2 = cloudy	1,406	18.96
	3 = rain	445	6
	4 = fog	107	1.44
	5 = sleet	15	0.2
	6 = snow	176	2.37
TEMP (temperature)	1 = <10 °C	2,074	27.97
	2 = 10–27 °C	3,624	48.88
	3 =>27 °C	1,716	23.15
NEAREST (intersecting in or near city)	1 = in city	4,244	57.24
	2 = near city	3,170	42.76
Crossing characteristics			
XSURFACE (crossing surface type)	1 = timber	2,049	27.64
	2 = asphalt	3,015	40.67
	3 = asphalt and flange	445	6
	4 = concrete	920	12.41
	5 = concrete and rubber	266	3.59
	6 = rubber	413	5.57
	7 = metal	3	0.04
	8 = unconsolidated	256	3.45
	9 = other	47	0.63
XBUCK (cross bucks)	1 = not present	2,348	31.67
	2 = present	5,066	68.33
FLASH (flashlight)	1 = not present	3,475	46.87
	2 = present	3,939	53.13
GATES (gates)	1 = not present	6,371	85.93
	2 = present	1,043	14.07

struck the vehicle while the remaining (21.36 %) were when vehicle struck the rail equipment. It is shown in the table that a majority (53.09 %) of vehicles involved in the vehicle-rail crashes are cars. It is also shown that the majority (71.01 %) of vehicle crashes had occurred in clear weather conditions.

The HRGC sites where crashes occurred are located in different development areas. As one can see from Table 1, 32.37 % of the crossings are located in open space areas,

Table 2 MNL model results

Parameter	Injury		Fatal	
	Estimate	*P* value	Estimate	*P* value
Intercept	−1.1553	<0.0001	−4.4843	<0.0001
VEHSPD (Ref: <40 km/h)				
40-72 km/h*	0.6457	<0.0001	0.7110	<0.0001
>72 km/h*	0.9211	<0.0001	1.6351	<0.0001
TYPVEH (Ref: auto)				
Truck	0.0581	0.6299	0.0846	0.6604
Truck-trailer*	−0.1967	0.0316	−1.5297	<0.0001
Pick-up truck*	0.1480	0.0766	0.0385	0.7808
Van	0.0756	0.6144	−0.2670	0.3401
Bus	0.7259	0.4470	−9.9575	0.9790
School bus	1.0507	0.2969	−10.0643	0.9820
TYPACC (Ref: vehicle struck rail equipment)				
Rail equipment struck vehicle*	−0.1107	0.1476	0.6935	<0.0001
TEMP (Ref: <10 °C)				
10–27 °C	0.1029	0.1654	0.0671	0.6081
>27 °C*	0.2520	0.0034	0.1148	0.4494
WEATHER (Ref: clear)				
Cloudy	−0.0399	0.6056	−0.0438	0.7463
Rain	−0.1611	0.2240	−0.4313	0.1130
Fog	0.0295	0.9021	−1.2110	0.1003
Sleet	0.4891	0.4086	−10.7328	0.9568
Snow*	−0.6097	0.0087	−0.6858	0.1285
TRNSPD (Ref: <40 km/h)				
40–72*	0.6274	<0.0001	1.7280	<0.0001
>72*	0.6433	<0.0001	2.7725	<0.0001
DRIVGEN (Ref: female)				
Male + missing*	0.3848	<0.0001	0.2965	0.0176
DEVELTYP(Ref: open space area)				
Residential	−0.1907	0.0231	−0.1882	0.1913
Commercial*	−0.3342	<0.0001	−0.3510	0.0171
Industrial*	−0.4128	<0.0001	−0.1197	0.5122
Institutional	−0.4649	0.0666	−0.5219	0.2897
XSURFACE(Ref: timber)				
Asphalt*	−0.2094	0.0043	−0.4813	0.0002
Asphalt and Flange	−0.1327	0.3229	−0.6683	0.0143
Concrete	0.0793	0.4405	0.0422	0.8002
Concrete and Rubber	0.0897	0.6240	0.5610	0.0428
Rubber	0.0745	0.6092	−0.3451	0.2467
Metal	−0.4770	0.7027	−10.1543	0.9825
Unconsolidated	−0.3027	0.0669	−0.1017	0.6871
Other	−0.2763	0.4752	−0.3334	0.6653
AADT(Ref: <10,000)				
10,000–20,000	−0.0882	0.4556	−0.4342	0.0698
20,000–30,000	−0.5348	0.0184	−0.8054	0.0755
>30,000	−0.2838	0.2595	−0.9880	0.0788
DRIVAGE (Ref: <25 years)				
25–60 years	0.0727	0.3548	0.2983	0.0452
>60 years*	0.2706	0.0069	1.2399	<0.0001

Number of observation = 7,414, $\rho^2 = 0.011$, χ^2 for likelihood ratio = 943.787, *P* value for $\chi^2 = 0.000$, Akike Information criteria (AIC) = 9667

Table 3 ORL model results

Parameter	Estimate	$Pr > \chi^2$
Intercept (1)	−3.2775	<0.0001
Intercept (2)	−1.1865	<0.0001
VEHSPD (Ref: < 40 km/h)		
40–72*	0.6051	<0.0001
>72*	0.9889	<0.0001
TYPVEH (Ref: auto)		
Truck	0.0717	0.5057
Truck-trailer*	−0.4956	<0.0001
Pick-up truck	0.1103	0.1437
Van	−0.0229	0.8689
Bus	0.4859	0.6081
School bus	0.5785	0.5562
TEMP (Ref: < 10 °C)		
10–27 °C	0.0834	0.2186
>27 °C	0.1973	0.0122
WEATHER (Ref: clear)		
Cloudy	−0.0566	0.4229
Rain	−0.2268	0.0683
Fog	−0.2184	0.3484
Sleet	0.0647	0.9139
Snow*	−0.6409	0.0028
TRNSPD (Ref: <40 km/h)		
40–72*	0.7720	<0.0001
>72*	1.2090	<.0001
DRIVAGE (Ref: <25 years)		
25–60 years	0.1334	0.0677
>60 years*	0.6123	<0.0001
DRIVGEN (Ref: female)		
Male + missing*	0.3431	<0.0001
DEVELTYP (Ref: open space area)		
Residential	−0.1800	0.0183
Commercial*	−0.3214	<0.0001
Industrial*	−0.3087	0.0006
Institutional	−0.4911	0.0382
XSURFACE (Ref: timber)		
Asphalt*	−0.2742	<0.0001

Table 3 continued

Parameter	Estimate	Pr > χ^2
Asphalt and flange	−0.2668	0.0340
Concrete	0.0669	0.4694
Concrete and rubber	0.2548	0.1120
Rubber	−0.0275	0.8406
Metal	−0.6138	0.6267
Unconsolidated	−0.2066	0.1573
Other	−0.2461	0.4905
AADT (Ref: <10,000)		
10,000–20,000	−0.1715	0.1214
20,000–30,000*	−0.5853	0.0057
>30,0000	−0.4343	0.0675

Likelihood ratio test $\chi^2 = 700.4685(37d.f.)$; P value is <0.0001

Score test for proportional odds assumption $\chi^2 = 206.5131(37d.f.)$; P value is <0.0001

Akaike information criterion (AIC) = 10478.994

21.51 % in residential areas, and 28.10 % in commercial areas. The rest are found in industrial and institutional development areas. The majority (74.99 %) of the HRGCs cross two-lane highways. Descriptive statistics of other variables are also shown in the table. As many variables as possible are considered in this study. Some of the continuous variables are converted into categorical variable and the multinomial logit model and ordered logit models are developed and compared to estimate the model parameters.

5 Results and discussion

Many variables obtained from the crossing inventory and crash data were used in developing the MNL and ORL models. During the final preferred model development process, some of the variables were found to be statistically insignificant and hence removed in a stepwise manner. PROC LOGISTIC procedure was applied with significance level being 0.1 to retain some of the variables.

Tables 2 through 5 present the MNL and ORL model results obtained from this study. In both models, three vehicle-rail crash severity levels (Fatal crashes, Injury crashes, and No Injury crashes) were considered as the dependent variable. In particular, in the MNL model, no injury crashes were considered the base case among the three crash severity levels. Therefore, coefficients estimated for the explanatory variables are values representing

Table 4 Marginal effects results for the MNL model

Variable	P (fatal)	P (injury)	P (no injury)
Vehicle speed (40–72 km/h)	0.028	0.135	−0.163
Vehicle speed(>72 km/h)	0.026	0.323	−0.349
Vehicle type (truck-trailer)	0.020	−0.316	0.296
Vehicle type (pick-up)	0.009	0.005	−0.014
Accident type (rail equipment struck vehicle)	−0.022	0.148	−0.125
Temperature (>27 °C)	0.014	0.019	−0.033
Weather (snow)	−0.026	−0.131	0.157
Weather (foggy)	0.028	−0.254	0.226
Train speed (40–72 km/h)	0.005	0.349	−0.353
Train speed(>72 km/h)	−0.017	0.567	−0.550
Vehicle driver gender (male)	0.019	0.054	−0.073
Development area type (residential)	−0.009	−0.035	0.044
Development area type (commercial)	−0.015	−0.066	0.081
Development area type (industrial)	−0.025	−0.016	0.041
Development area type (institutional)	−0.020	−0.099	0.119
HRGC surface type (asphalt)	−0.004	−0.096	0.100
HRGC surface type (unconsolidated)	−0.018	−0.015	0.033
HRGC surface type (asphalt and flange)	0.006	−0.137	0.132
HRGC surface type (concrete and rubber)	−0.006	0.116	−0.110
Traffic volume (AADT 10,000–20,000)	0.003	−0.089	0.086
Traffic volume (AADT 20,000–30,000)	−0.018	−0.157	0.176
Traffic volume (AADT >30,000)	0.002	−0.201	0.199
Vehicle driver age (25–60 years)	−0.002	0.061	−0.059
Vehicle driver age (>60 years)	−0.009	0.254	−0.245

the relative effect of contributing factors on fatal or injury crashes compared to no injury crashes. Positive estimates in the models indicate that the chance of injury or fatal crash increases as the value of the independent variables increases. On the other hand, the interpretation of the coefficients in the ORL model is different and can be presented as follows: A positive coefficient in the model indicates that an increase in the value of a variable will increase the probability of the highest severity level (fatal) and decrease the probability of the lowest severity level (no injury). On the other hand, a negative coefficient indicates that a decrease in the variable will increase the probability of the highest severity level and decrease the probability of the lowest severity level. For the intermediate severity level (injury), an increase in the value of a variable may decrease or increase the probability of it occurring.

As shown in Tables 2 and 3, some of the variables are not statistically significant. However, for the convenience of interpretation, those variables were still retained in the model if at least one of variables/factors in the same parameter category was significant in at least one of the models (injury and/or fatality), though this may actually induce reduction in efficiency of the model. Furthermore, a 90 % confidence level was considered instead of 95 %.

It is important to note that the assessment and comparison of the two models cannot be performed simply based on the estimated coefficients of the models. Marginal effects of the variables on the probability of severity levels are computed for the two models in Tables 4 and 5 and used for comparison purpose. The positive sign in estimated marginal effect indicates that the probability of a given crash severity level increases when the variable changes and the converse is true for a negative sign. And the value of the number indicates the magnitude of shift in the probability.

The shifting direction of the probability in the two models was used for comparison of the impacts of each variable on the probability of injury severity outcomes as shown in Table 6. As the results indicate, most variables are consistent which include the variable crash circumstance for the case of intermediate severity level (injury). Empty cell indicates that the variable is not significant even at the 90 % confidence level.

Some of the variables in the ORL model, including pick-up vehicles, crash circumstances when rail equipment struck vehicle, Foggy weather, unconsolidated, and concrete & rubber surface type and traffic volume (AADT of 10,000–20,000) are found to be statistically insignificant while they all are statistically significant in the case of MNL model. On the other hand, rainy weather condition was found to be statistically significant in the MNL model. In addition, some of the other variables are found inconsistent at both fatality and no injury severity levels, all of which are highlighted with red color and "+" or "−" signs. However, a majority of the variables have shown similar effects on the probability of the three different severity levels. The Akaike information criterion (AIC) of the two models is 9,667 and 10,479 for the MNL and ORL

Table 5 Marginal effects results for the ORL model

Variable	*P* (fatal)	*P* (injury)	*P* (no injury)
Indicator for vehicle speed is category 2 (40–72 km/h)	0.027	0.009	−0.036
Indicator for vehicle speed is category 3 (>72 km/h)	0.045	0.015	−0.059
Indicator for vehicle type truck-trailer	−0.022	−0.007	0.030
Indicator for higher temperature (>27 °C)	0.009	0.003	−0.012
Indicator for rainy weather	−0.010	−0.003	0.014
Indicator for snow weather	−0.029	−0.010	0.038
Indicator for train speed is category 2 (40–72 km/h)	0.035	0.012	−0.046
Indicator for train speed is category 3 (>72 km/h)	0.054	0.018	−0.073
Indicator for vehicle driver age 25–60 years	0.006	0.002	−0.008
Indicator for vehicle driver age >60 years	0.028	0.009	−0.037
Indicator for vehicle driver gender male	0.015	0.005	−0.021
Indicator for residential development area type	−0.008	−0.003	0.011
Indicator for commercial development area type	−0.014	−0.005	0.019
Indicator for industrial development area type	−0.014	−0.005	0.019
Indicator for institutional development area type	−0.022	−0.007	0.029
Indicator for HRGC asphalt surface type	−0.012	−0.004	0.016
Indicator for HRGC asphalt and flange surface type	−0.012	−0.004	0.016
Indicator for traffic volume (AADT of 20,000–30,000)	−0.026	−0.009	0.035
Indicator for traffic volume (AADT of >30,000)	−0.020	−0.007	0.026

Table 6 Comparison of Marginal Effects on Variables for ORL and MNL Models

Variable	Fatal		Injury		No injury	
	ORL	MNL	ORL	MNL	ORL	MNL
Indicator for vehicle speed is category 2 (40-72km/hour)	+	+	+	+	-	-
Indicator for vehicle speed is category 3 (>72km/hour)	+	+	+	+	-	-
Indicator for vehicle type truck-trailer	-	+	-	-	+	+
Vehicle type (Pick-up)		+		+		-
Crash circumstance (rail equipment struck vehicle)		-		+		-
Indicator for higher temperature (>27ºC)	+	+	+	+	-	-
Indicator for rainy weather	-		-		+	
Indicator for snow weather	-	-	-	-	+	+
Weather (foggy)		+		-		+
Indicator for train speed is category 2 (40-72km/hour)	+	+	+	+	-	-
Indicator for train speed is category 3 (>72km/hour)	+	-	+	+	-	-
Indicator for vehicle driver age 25-60 years	+	-	+	+	-	-
Indicator for vehicle driver age >60 years	+	-	+	+	-	-
Indicator for vehicle driver gender male	+	+	+	+	-	-
Indicator for residential development area type	-	-	-	-	+	+
Indicator for commercial development area type	-	-	-	-	+	+
Indicator for industrial development area type	-	-	-	-	+	+
Indicator for institutional development area type	-	-	-	-	+	+
HRGC surface type (Unconsolidated)		-		-		+
HRGC surface type (Concrete and rubber)		-		+		-
Indicator for HRGC asphalt surface type	-	-	-	-	+	+
Indicator for HRGC asphalt and flange surface type	-	+	-	-	+	+
Traffic volume (AADT 10,000-20,000)		+		-		+
Indicator for traffic volume (AADT of 20,000-30,000)	-	-	-	-	+	+
Indicator for traffic volume (AADT of >30,000)	-	+	-	-	+	+

model, respectively. This indicates that the MNL model is better than the ORL model in predicting vehicle crash severity at HRGCs in this paper.

6 Conclusion

Comparison between the MNL and ORL model in predicting the vehicle crash severity on HRGCs was conducted using the USDOT public crossing sites data. The three vehicle crash severity levels, fatality, injury, and no injury, were considered as dependent variable. Train characteristics, environmental characteristics, types of development areas, highway-rail crossing characteristics, highway traffic characteristics, vehicle driver characteristics, and vehicle characteristics were the explanatory variables used in predicting the vehicle crash severity levels. The analysis was conducted using SAS PROC LOGISTIC procedure.

As discussed in the result part of this paper, in the ORL model, some variables were found to be statistically significant while they were not in the MNL model and vice versa. A majority of the variables have shown similar effects on the probability of the three different severity levels. In addition, based on the AIC, it was found that the MNL model is better than the ORL model in predicting the vehicle crash severity levels on HRGCs in this study. Therefore, the researcher recommends the MNL model to be applied rather than the

ORL model in predicting severity levels of vehicle-rail crashes on highway-rail at-grade crossings.

References

1. National Highway Traffic Safety Administration (2012) US Department of transportation Traffic Safety Facts Research Note. 2011 Motor vehicle crashes: overview. http://www-nrd.nhtsa.dot.gov/Pubs/811701.pdf. Accessed 05 June 2015
2. Federal Railroad Administration (2012) US Department of Transportation. Railroad safety statistics, 2012 preliminary annual report. http://safetydata.fra.dot.gov/OfficeofSafety/publicsite/Prelim.aspx. Accessed 05 June 2015
3. Savolainen P, Mannering F, Lord D, Quddus M (2011) The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. Accid Anal Prev 43:1666–1676
4. Long JS (1997) Regression models for categorical and limited dependent variables. Sage Publications, Thousand Oaks
5. Fan W, Haile EW (2014). Analysis of severity of vehicle crashes at highway-rail grade crossings: multinomial logit modeling. Paper No. 14-0588. Proceedings of Transportation Research Board 93rd Annual Meeting, Washington D.C
6. Mercier C, Shelley M, Adkins G, Mercier J (1999) Age and gender as predictors of injury severity in broadside and angle vehicular collisions. Paper No. 99-0607. Transportation Research Record, pp 50–61
7. Dissanayake S, Lu J (2002) Analysis of severity of young driver crashes: sequential binary logistic regression modeling. Paper No. 02-2302. Transportation Research Record, pp 108–114
8. Duncan C, Khattak A, Council F (1998) Applying the ordered probit model to injury severity in truck–passenger car rear-end collisions. Paper No. 98-1237. Transportation Research Record, pp 63–71
9. Donnell ET, Mason JM (2004) Predicting the severity of median-related crashes in pennsylvania by using logistic regression. Paper No. 1897. Transportation Research Record: Journal of the Transportation Research Board, pp 55–63
10. Renski H, Khattak AJ, Council FM (1999) Effect of speed limit increases on crash injury severity analysis of single-vehicle crashes on North Carolina interstate highways. Paper No. 99-0975 Transportation Research Record, pp 100–108
11. Huang HF, Stewart JR, Zegeer CV (2002) Evaluation of lane reduction "road diet" measures on crashes and injuries. Paper No. 02-2955. Transportation Research Record, pp 80–90
12. Khattak AJ (2001) Injury severity in multivehicle rear-end crashes. Paper No. 01-3466. Transportation Research Record, pp 59–68
13. Mercier C, Shelley M, Rimkus J, Mercier J (1997) Age and gender as predictors of injury severity in head-on highway vehicular collisions. Paper No. 97-0535. Transportation Research Record, pp 37–46
14. Chira-Chavala T, Coifman B, Porter C, Hansen M (1996) Light rail accident involvement and severity. Transp Res Rec 1521:147–155
15. Chen W, Jovanis P (2000) Method for identifying factors contributing to driver-injury severity in traffic crashes. Paper No. 00-1707. Transportation Research Record, pp 1–9
16. Khattak A, Pawlovich M, Souleyrette R, Hallmark S (2002) Factors related to more severe older driver traffic crash injuries. J Transp Eng 128:243–249
17. Xie Y, Zhang Y, Faming L (2009) Crash injury severity analysis using bayesian ordered probit models. J Transp Eng 135:18–25
18. Zhao S, Khattak A (2015) Motor vehicle drivers' injuries in train-Motor vehicle crashes. Accid Anal Prev 74:162–168
19. SAS/STAT User's Guide, Version 9.2. (2008) SAS Publishing, Cary